Báo cáo bài tập môn Xử lý Ngôn ngữ tự nhiên

Sinh viên: Nguyễn Minh Quang 19020405
Nguyễn Như Ngọc 19020385

Nội dung bài tập

Bài 1. Sử dụng độ đo Cosine để đo độ tương đồng của cặp từ đã cho

Bài 2. Tìm k tự có độ tương tự cao nhất đối với một từ w cho trước

Bài 3. Xây dựng mô hình phân loại từ đồng nghĩa trái nghĩa

Nội dung báo cáo:

Trong quá trình khảo sát bộ dữ liệu, trong các bộ dữ liệu có xuất hiện những từ vựng nằm ngoài bộ word2vec tiếng Việt được cho sẵn. Hiện tại nhóm chưa có phương án để dự đoán vector của các từ này cho nên các từ vựng này (cặp từ vựng tương ứng) sẽ được loại bỏ trong quá trình làm bài.

Code và dữ liệu

Bài 1

Với bộ **ViSim-400** được sử dụng để tính độ đo tương tự Cosine, ta lọc được 344 cặp từ có biểu diễn vector trong bộ word2vec W2V_150 đã cho.

Word1	Word2	POS	Sim1	Sim2	STD	sim-cosine
biến	ngặp	V	3.13	5.22	0.72	-0.004912
nhà_thi_đấu	nhà	N	3.07	5.12	1.18	0.082523
động	tĩnh	V	0.60	1.00	0.95	0.277086
khuyết	ưu	N	0.20	0.33	0.40	0.176799
thủ_pháp	biện_pháp	N	4.13	6.88	1.26	0.402366
…	…	…	…	…	…	…

Đánh giá độ tương đồng giữa kết quả Sim1 và sim-cosine là:

Pearson correlation coefficient:

correlation=0.4468430791767022, p-value=2.7457782359140626e-18

Spearman's rank correlation coefficient:

correlation=0.4077422929392862, p-value=3.2726552283981985e-15

Bài 2

Dựa trên độ đo Consine được xây dựng từ bài 1, để tìm k từ có độ tương đồng cao nhất với một từ w cho trước:

- Quét và tính toàn bộ độ giá trị cosine-sim của w với các từ còn lại trong bộ từ vựng
- Sắp xếp theo giá trị cosine-sim giảm dần
- Lấy ra k từ đầu tiên trong dữ liệu cosine-sim

Ví dụ:

Với từ "**tượng_đài**" là một từ có trong bộ từ vựng ta tìm được 5 từ tương tự nhất là

'đền_thờ'	0.5623567247142175
'tháp_chuông'	0.5443984164585618
bia_tưởng_niệm'	0.5406205813189373
'giáo_đường'	0.5329588054503885
'lăng_mộ'	0.5287220564708057

Với từ "**tương** lai" là một từ có trong bộ từ vựng ta tìm được 5 từ tương tự nhất là

'tương_lại'	0.6938872953960925
'tương_lai_gần'	0.6184742365234566
'viễn_cảnh'	0.5908744782640388
'trong_tương_lai'	0.5701547617009627
'trung_hạn'	0.4688059746341524

Như vậy với một số ví dụ, chúng ta có thể thấy sử dụng độ đo tương tự cosine có thể tính toán được độ tương tư giữa các từ với một mức độ chính xác nhất định.

Bài 3

Sử dụng mô hình **multiple-perceptron** để dự đoán đồng nghĩa, trái nghĩa

Bộ dữ liệu dùng để huấn luyện mô hình là **antonym-synonym** gồm có 2000 cặp từ trái nghĩa, 11562 cặp từ đồng nghĩa

Bộ dữ liệu dùng để đánh giá mô hình là **ViCon-400** gồm có

- 400 cặp Danh từ đồng nghĩa / trái nghĩa
- 400 cặp Động từ đồng nghĩa / trái nghĩa
- 400 cặp Tính từ đồng nghĩa / trái nghĩa

Trước hết ta tiến hành xử lý dữ liệu huấn luyện và dữ liệu đánh giá. Sau khi loại bỏ các từ / cặp từ không nằm trong bộ từ vựng, các cặp từ lặp lại ta thu được

- Bộ dữ liệu huấn luyện gồm 7350 cặp từ
- Bộ dữ liệu đánh giá gồm 835 cặp từ (đã loại bỏ các cặp từ đã xuất hiện trong tập huấn luyện)

Mô hình dùng được huấn luyện ở có kiến trúc như sau:

Layer (type)	Output Shape	Param #
Linear-1	[-1, 150, 500]	75,500
Relu-1	[-1, 150, 500]	
Linear-2	[-1, 150, 500]	250,500
Tanh-1	[-1, 150, 500]	
Linear-3	[-1, 150, 500]	250,500
Relu-2	[-1, 150, 500]	
Linear-4	[-1, 150, 500]	250,500
Tanh-2	[-1, 150, 500]	
Linear-5	[-1, 2]	1,004
Soft-max-1	[-1, 2]	

Các tham số huấn luyện

Optimizer = SGD(learning_rate=0.0001, momentum=0.9)

Epochs = 20

Batch_size = 32

Kết quả dự đoán trên tập đánh giá (đã được loại từ ngoài tập từ vựng và từ trùng với tập huấn luyện)

	Precision	Recall	F1-score	Support
ANT	1.00	0.91	0.95	466
SYN	0.90	1.00	0.95	369
accuracy			0.95	835
macro avg	0.95	0.96	0.95	835
weighted avg	0.96	0.95	0.95	835